

# Diffusion and Superposition Distances for Signals Supported on Networks

Santiago Segarra, Weiyu Huang and Alejandro Ribeiro

**Abstract**—We introduce the diffusion and superposition distances as two metrics to compare signals supported in the nodes of a network. Both metrics consider the given vectors as initial temperature distributions and diffuse heat through the edges of the graph. The similarity between the given vectors is determined by the similarity of the respective diffusion profiles. The superposition distance computes the instantaneous difference between the diffused signals and integrates the difference over time. The diffusion distance determines a distance between the integrals of the diffused signals. We prove that both distances define valid metrics and that they are stable to perturbations in the underlying network. We utilize numerical experiments to illustrate their utility in classifying signals in a synthetic network as well as in classifying ovarian cancer histologies using gene mutation profiles of different patients. We also reinterpret diffusion as a transformation of interrelated feature spaces and use it as preprocessing tool for learning. We use diffusion to increase the accuracy of handwritten digit classification.

## I. INTRODUCTION

Networks, or graphs, are data structures that encode relationships between elements of a group and which, for this reason, play an important role in many disparate disciplines such as biology [1], [2] and sociology [3], [4] where relationships between, say, genes, species or humans, are central. Often, networks have intrinsic value and are themselves the object of study. This is the case, e.g., when we are interested in distributed and decentralized algorithms in which agents iterate through actions that use information available either locally or at adjacent nodes to accomplish some sort of global outcome [5]–[7]. Equally often, the network defines an underlying notion of proximity, but the object of interest is a signal defined on top of the graph. This is the matter addressed in the field of graph signal processing, where the notions of frequency and linear filtering are extended to signals supported on graphs [8]–[12]. Examples of network-supported signals include gene expression patterns defined on top of gene networks [13] and brain activity signals supported on top of brain connectivity networks [14]. Indeed, one of the principal uses of networks of gene interactions is to determine how a change in the expression of a gene, or a group of genes, cascades through the network and alters the expression of other genes. Likewise, a brain connectivity network specifies relationships between areas of the brain, but it is the pattern of activation of these regions that determines the mental state of the subject.

In this paper we consider signals supported on graphs and address the challenge of defining a notion of distance between these signals that incorporates the structure of the underlying network. We want these distances to be such that two signals are deemed close if they are themselves close – in the examples in the previous paragraph we have gene expression or brain activation patterns that are similar –, or if they have similar values in adjacent or nearby nodes – the expressed genes or the active areas of the brain are not similar but they effect similar changes in the gene network or represent activation of closely connected areas of the brain. We define here the diffusion and superposition distances and argue that they inherit this functionality through their connection to diffusion processes.

Diffusion processes draw their inspiration from the diffusion of heat through continuous matter [15], [16]. The linear differential equation that models heat diffusion can be extended to encompass dynamics through discrete structures such as graphs or networks [17]–[21]. In the particular case of graphs, every node is interpreted as containing an amount of heat which flows from hot to cold nodes. The flow of heat is through the edges of the graph and such that the rate at which heat diffuses is proportional to a weight that defines the proximity between the nodes adjacent to the edge. Diffusion processes in graphs are often used in engineering and science because they reach isothermal configurations in steady state. Driving the network to an isothermal equilibrium is tantamount to achieving a consensus action [22], [23], which, in turn, is useful in, e.g., problems in formation control [24] and flocking [25], as well as an important modeling tool in situations such as the propagation of opinions in social networks [26]–[28].

In this paper we do not exploit the asymptotic, but rather the transient behavior of diffusion processes. We regard the given vectors as initial heat configurations that generate different diffused heat profiles over time. The diffusion and superposition distances between the given vectors are defined as the difference between these heat profiles integrated over time. The superposition distance compares the instantaneous difference between the two evolving heat maps and integrates this difference over time. The diffusion metric integrates each of the heat profiles over time and evaluates the norm of the difference between the two integrals. Both of these distances yield small values when the diffusion profiles are similar. This happens if the given vectors themselves are close or if they have similar values at nodes that are linked by edges with high similarity values.

Work in this paper is supported by NSF CCF-1217963. The authors are with the Department of Electrical and Systems Engineering, University of Pennsylvania, 200 South 33rd Street, Philadelphia, PA 19104. Email: {ssegarra, whuang, aribeiro}@seas.upenn.edu.

## A. Contributions and summary

Besides the definition of the superposition and diffusion distances, the contributions of this paper are: (i) To prove that the superposition and diffusion distances are valid metrics in the space of vectors supported on a given graph. (ii) To show that both distances are well behaved with respect to small perturbations in the underlying network. (iii) To illustrate their ability to identify vectors that are similar only after the network structure is accounted for. (iv) To demonstrate their value in two practical scenarios; the classification of ovarian cancer types from gene mutation profiles and the classification of handwritten arabic digits.

We begin the paper with a brief introduction of basic concepts in graph theory and metric geometry followed by a formal description of diffusion dynamics in networks (Section II). This preliminary discussion provides the necessary elements for a formal definition of the superposition and diffusion distances. In Section III we define the superposition distance between two signals with respect to a given graph and a given input norm. To determine this distance the signals are diffused in the graph, the input norm of their difference is computed for all times, and the result is discounted by an exponential factor and integrated over time. We show that the superposition distance is a valid metric between vectors supported in the node set of a graph.

The diffusion distance with respect to a given graph and a given input norm is introduced in Section IV as an alternative way of measuring the distance between two signals in a graph. In this case the diffused signals are also exponentially discounted and integrated over time but the input norm is taken after time integration. The diffusion distance is shown to also be a valid metric in the space of signals supported on a given graph and is further shown to provide a lower bound for the superposition distance. Different from the superposition distance, the diffusion distance can be reduced to a closed form expression with a computational cost that is dominated by a matrix inversion. The superposition distance requires numerical integration of the time integral of the norm of a matrix exponential.

We further address stability with respect to uncertainty in the specification of the network (Section V). Specifically, we prove that when the input norm is either the 1-norm, the 2-norm, or the infinity-norm a small perturbation in the underlying network transports linearly to a small perturbation in the values of the superposition and diffusion distances. In Section VI we demonstrate that the diffusion and superposition distances can be applied to classify signals in graphs with better accuracy than comparisons that utilize traditional vector distances. We illustrate the differences using synthetic data (Section VI-A) and establish the practical advantages through the classification of ovarian cancer histologies from gene mutation profiles of different patients (Section VI-B). In Section VII, we reinterpret diffusion as a method for data preprocessing in learning for cases where interrelations exist across features in the feature space. We show the benefit of this data preprocessing through the classification of handwritten digits. We offer concluding remarks in Section VIII.

## II. PRELIMINARIES

We consider networks that are weighted, undirected, and symmetric. Formally, we define a network as a graph  $G = (V, E, W)$ , where  $V = \{1, \dots, n\}$  is a finite set of  $n$  nodes or vertices,  $E \subseteq V \times V$  is a set of edges defined as ordered pairs  $(i, j)$ , and  $W : E \rightarrow \mathbb{R}_{++}$  is a set of strictly positive weights  $w_{ij} > 0$  associated with each edge  $(i, j)$ . Since the graph is undirected, we must have that the edge  $(i, j) \in E$  if and only if  $(j, i) \in E$ . Since the graph is also symmetric, we must have  $w_{ij} = w_{ji}$  for all  $(i, j) \in E$ . The edge  $(i, j)$  represents the existence of a relationship between  $i$  and  $j$  and we say that  $i$  and  $j$  are adjacent or neighboring. The weight  $w_{ij} = w_{ji}$  represents the strength of the relationship, or, equivalently, the proximity or similarity between  $i$  and  $j$ . Larger edge weights are interpreted as higher similarity between the border nodes. The graphs considered here do not contain self loops, i.e.,  $(x, x) \notin E$  for any  $x \in V$ .

We consider the usual definitions of the adjacency, Laplacian, and degree matrices for the weighted graph  $G = (V, E, W)$ ; see e.g. [29, Chapter 1]. The adjacency matrix  $A \in \mathbb{R}_+^{n \times n}$  is such that  $A_{ij} = w_{ij}$  whenever  $i$  and  $j$  are adjacent, i.e., whenever  $(i, j) \in E$  and such that for  $(i, j) \notin E$  we have  $A_{ij} = 0$ . The degree matrix  $D \in \mathbb{R}_+^{n \times n}$  is a diagonal matrix such that the  $i$ -th diagonal element  $D_{ii} = \sum_j w_{ij}$  contains the sum of all the weights out of node  $i$ . The Laplacian matrix is defined as the difference  $L := D - A \in \mathbb{R}^{n \times n}$ . Since  $D$  is diagonal and the diagonal of  $A$  is null – because  $G$  does not have self loops – the components of the Laplacian matrix are explicitly given by

$$L_{ij} := \begin{cases} -A_{ij} & \text{if } i \neq j, \\ \sum_{k=1}^n A_{ik} & \text{if } i = j. \end{cases} \quad (1)$$

Observe that the Laplacian is positive semidefinite because it is diagonally dominant with positive diagonal elements.

### A. Metrics and norms

Our goal in this paper is to define a metric to compare vectors defined on top of a graph. For reference, recall that for a given space  $X$ , a metric  $d : X \times X \rightarrow \mathbb{R}_+$  is a function from pairs of elements in  $X$  to the nonnegative reals satisfying the following three properties for every  $x, y, z \in X$ :

*Symmetry:*  $d(x, y) = d(y, x)$ .

*Identity:*  $d(x, y) = 0$  if and only if  $x = y$ .

*Triangle inequality:*  $d(x, y) \leq d(x, z) + d(z, y)$ .

A closely related definition is that of a norm. In this case we need to have a given vector space  $Y$  and consider elements  $v \in Y$ . A norm  $\|\cdot\|$  is a function  $\|\cdot\| : Y \rightarrow \mathbb{R}_+$  from  $Y$  to the nonnegative reals such that, for all vectors  $v, w \in Y$  and scalar constant  $\beta$ , it satisfies:

*Positiveness:*  $\|v\| \geq 0$  with equality if and only if  $v = \vec{0}$ .

*Positive homogeneity:*  $\|\beta w\| = |\beta| \|w\|$ .

*Subadditivity:*  $\|v + w\| \leq \|v\| + \|w\|$ .

Norms are more stringent than metrics because they require the existence of a null element with null norm. However,

whenever a norm is defined on a vector space  $Y$  it induces a distance in the same space as we formally state next [30, Chapter 1].

**Lemma 1** *Given any norm  $\|\cdot\|$  on some vector space  $Y$ , the function  $d : Y \times Y \rightarrow \mathbb{R}_+$  defined as  $d(r, s) := \|r - s\|$  for all pairs  $r, s \in Y$  is a metric.*

In some of our proofs we encounter norms induced in the vector space of matrices  $\mathbb{R}^{n \times n}$  by norms defined in the vector space  $\mathbb{R}^n$ . For a given vector norm  $\|\cdot\| : \mathbb{R}^n \rightarrow \mathbb{R}_+$  the induced matrix norm  $\|\cdot\| : \mathbb{R}^{n \times n} \rightarrow \mathbb{R}_+$  is defined as

$$\|A\| := \sup_{\|x\|=1} \|Ax\|. \quad (2)$$

I.e., the norm of matrix  $A$  is equal to the maximum vector norm achievable when multiplying  $A$  by a vector with unit norm. Apart from satisfying the three requirements in the definition of norms, induced matrix norms are compatible and submultiplicative [31, Section 2.3]. That they are submultiplicative means that for any given pair of matrices  $A, B \in \mathbb{R}^{n \times n}$  the norm of the product does not exceed the product of the norms,

$$\|AB\| \leq \|A\| \|B\|. \quad (3)$$

That they are compatible means that for any vector  $x \in \mathbb{R}^n$  and matrix  $A \in \mathbb{R}^{n \times n}$  it holds,

$$\|Ax\| \leq \|A\| \|x\|. \quad (4)$$

I.e., the vector norm of the product  $Ax$  does not exceed the product of the norms of the vector  $x$  and the induced norm of the matrix  $A$ .

### B. Diffusion dynamics

Consider an arbitrary graph  $G = (V, E, W)$  with Laplacian matrix  $L$  and a vector  $r = [r_1, \dots, r_n]^T \in \mathbb{R}^n$  where the component  $r_i$  of  $r$  corresponds to the node  $i$  of  $G$ . For a given constant  $\alpha > 0$ , define the time-varying vector  $r(t) \in \mathbb{R}^n$  as the solution of the linear differential equation

$$\frac{dr(t)}{dt} = -\alpha L r(t), \quad r(0) = r. \quad (5)$$

The differential equation in (5) represents heat diffusion on the graph  $G$  because  $-L$  can be shown to be the discrete approximation of the continuous Laplacian operator used to describe the diffusion of heat in physical space [17]. The given vector  $r = r(0)$  specifies the initial temperature distribution and  $r(t)$  represents the temperature distribution at time  $t$ . The constant  $\alpha$  is the thermal conductivity and controls the heat diffusion rate. Larger  $\alpha$  results in faster changing  $r(t)$ . The solution of (5) is given by the matrix exponential,

$$r(t) = e^{-\alpha L t} r, \quad (6)$$

as can be verified by direct substitution of  $r(t) = e^{-\alpha L t} r$  in (5). The expression in (6) allows us to compute the temperature distribution at any point in time given the initial heat configuration  $r$  and the structure of the underlying network through its Laplacian  $L$ . Notice that as time grows,  $r(t)$  settles to an

isothermal equilibrium – all nodes have the same temperature – if the graph is connected.

It is instructive to rewrite (5) componentwise. If we focus on the variation of the  $i$ -th component of  $r(t)$  and use the definition of  $L$  in (1) to replace  $L_{ik} = -A_{ik}$  and  $L_{ii} = \sum_{k=1}^n A_{ik}$ , it follows that (5) implies

$$\frac{dr_i(t)}{dt} = \sum_{k=1}^n \alpha A_{ik} (r_k(t) - r_i(t)). \quad (7)$$

Further recalling that  $A_{ik} = 0$  if  $i$  and  $k$  are not adjacent and that  $A_{ik} = w_{ik}$  otherwise, we see that the sum in (7) entails multiplying each of the differences  $r_k(t) - r_i(t)$  between adjacent nodes by the corresponding proximities  $w_{ik}$ . Thus, (7) is describing the flow of heat through edges of the graph. The flow of heat on an edge grows proportionally with the temperature differential  $r_k(t) - r_i(t)$ , but also with the proximity  $w_{ik}$ . Nodes with large proximity tend to equalize their temperatures faster, other things being equal. In particular, two initial vectors  $r(0) = r$  and  $s(0) = s$  result in similar temperature distributions across time if they are themselves similar – all  $r_i$  and  $s_i$  components are close –, or if they have similar initial levels at nodes with large proximity – each component  $r_i$  may not be similar to  $s_i$  itself but similar to the component  $s_j$  of a neighboring node for which the edge weight  $w_{ij}$  is large. This latter fact suggests that the diffused vectors  $r(t)$  and  $s(t)$  define a notion of proximity between  $r$  and  $s$  associated with the underlying graph structure. We exploit this observation to define distances between signals supported on graphs in the following two sections.

### III. SUPERPOSITION DISTANCE

Given an arbitrary graph  $G = (V, E, W)$  with Laplacian matrix  $L$ , an input vector norm  $\|\cdot\|$ , and two signals  $r, s \in \mathbb{R}^n$  defined in the node space  $V$ , we define the superposition distance  $d_{\text{sps}}^L(r, s)$  between  $r$  and  $s$  as

$$d_{\text{sps}}^L(r, s) := \int_0^{+\infty} e^{-t} \|e^{-\alpha L t} (r - s)\| dt, \quad (8)$$

where  $\alpha > 0$  corresponds to the diffusion constant in (5). As we mentioned in the discussion following (7), the distance  $d_{\text{sps}}^L(r, s)$  defines a similarity between  $r$  and  $s$  that incorporates the underlying network structure. Indeed, notice that the term inside the input norm corresponds to the difference  $r(t) - s(t)$  between the vectors that solve (5) for initial conditions  $r$  and  $s$  [cf. (6)]. This means that we are looking at the difference between the temperatures  $r(t)$  and  $s(t)$  at time  $t$ , which we then multiply by the dampening factor  $e^{-t}$  and integrate over all times. These temperatures are similar if  $r$  and  $s$  are similar, or, if  $r$  and  $s$  have similar values at similar nodes. The dampening factor gives more relative importance to the differences between  $r(t)$  and  $s(t)$  for early times. This is necessary because after prolonged diffusion times the network settles into an isothermal equilibrium and the structural differences between  $r$  and  $s$  are lost.

Exploiting the same interpretation, we can define the superposition norm of a vector  $v \in \mathbb{R}^n$  for a given graph with

Laplacian matrix  $L$  and a given input norm  $\|\cdot\|$  as

$$\|v\|_{\text{sps}}^L := \int_0^{+\infty} e^{-t} \|e^{-\alpha L t} v\| dt. \quad (9)$$

Although we are referring to  $d_{\text{sps}}^L(r, s)$  as the superposition distance between  $r$  and  $s$  and  $\|v\|_{\text{sps}}^L$  as the superposition norm of  $v$  we have not proven that they indeed are valid definitions of distance and norm functions. As it turns out, they are. We begin by showing that  $\|\cdot\|_{\text{sps}}^L$  is a valid norm as we claim in the following proposition.

**Proposition 1** *The function  $\|\cdot\|_{\text{sps}}^L$  in (9) is a valid norm on  $\mathbb{R}^n$  for every Laplacian  $L$  and every input norm  $\|\cdot\|$ .*

**Proof:** As stated in Section II, we need to show positiveness, positive homogeneity and subadditivity of  $\|\cdot\|_{\text{sps}}^L$ . To show positive homogeneity, utilize the positive homogeneity of the input norm and the linearity of integrals to see that for every vector  $v \in \mathbb{R}^n$  and scalar  $\beta$ , it holds

$$\begin{aligned} \|\beta v\|_{\text{sps}}^L &= \int_0^{+\infty} e^{-t} \|e^{-\alpha L t} \beta v\| dt \\ &= |\beta| \int_0^{+\infty} e^{-t} \|e^{-\alpha L t} v\| dt \\ &= |\beta| \|v\|_{\text{sps}}^L. \end{aligned} \quad (10)$$

In order to show subadditivity, pick arbitrary vectors  $v, w \in \mathbb{R}^n$  and use the subadditivity of the input norm  $\|\cdot\|$  and the linearity of integrals to see that

$$\begin{aligned} \|v + w\|_{\text{sps}}^L &= \int_0^{+\infty} e^{-t} \|e^{-\alpha L t} (v + w)\| dt \\ &\leq \int_0^{+\infty} e^{-t} (\|e^{-\alpha L t} v\| + \|e^{-\alpha L t} w\|) dt \\ &= \|v\|_{\text{sps}}^L + \|w\|_{\text{sps}}^L, \end{aligned} \quad (11)$$

To show positiveness, first observe that for every  $v \in \mathbb{R}^n$  we have that  $\|v\|_{\text{sps}}^L \geq 0$  since for every time  $t$  the argument of the integral in the definition (9) is the product of two nonnegative terms, an exponential and a norm which itself satisfies the positiveness property. The fact that  $\|\vec{0}\|_{\text{sps}}^L = 0$  is an immediate consequence of the definition (9). Hence, we are only left to show that  $\|v\|_{\text{sps}}^L \neq 0$  for  $v \neq 0$ . To show this, it suffices to prove that the argument of the integral in (9) is strictly positive for every time  $t$  which is implied by the fact that the matrix  $e^{-\alpha L t}$  is strictly positive definite for every  $t$ . To see why this is true, notice that  $-\alpha L t$  is a real symmetric matrix, thus, it is diagonalizable and has real eigenvalues. Consequently, the eigenvalues of  $e^{-\alpha L t}$  are the exponentials of the eigenvalues of  $-\alpha L t$  which are strictly positive. ■

If the superposition norm is a valid norm as shown by Proposition 1 it induces a valid metric as per the construction in Lemma 1. This induced metric is the superposition distance defined in (8) as we show in the following corollary.

**Corollary 1** *The function  $d_{\text{sps}}^L$  in (8) is a valid metric on  $\mathbb{R}^n$  for every Laplacian  $L$  and every input norm  $\|\cdot\|$ .*

**Proof:** Since  $d_{\text{sps}}^L(r, s) = \|r - s\|_{\text{sps}}^L$  for all vectors  $r, s \in \mathbb{R}^n$  and  $\|\cdot\|_{\text{sps}}^L$  is a well-defined norm [cf. Proposition 1], Lemma 1 implies that  $d_{\text{sps}}^L$  is a metric on  $\mathbb{R}^n$ . ■

The distance  $d_{\text{sps}}^L$  incorporates the network structure to compare two signals  $r$  and  $s$  supported in a graph with Laplacian  $L$ . As a particular case the edge set  $E$  of the underlying graph  $G$  may be empty. In this case, the Laplacian  $L = \mathbf{0}$  is identically null and we obtain from (8) that  $d_{\text{sps}}^{\mathbf{0}}(r, s) = \|r - s\|$ . This is consistent with the fact that when no edges are present, the network structure adds no information to aid in the comparison of  $r$  and  $s$  and the superposition distance reduces to the standard distance induced by the input norm.

The computational cost of evaluating the superposition distance is significant in general. To evaluate  $d_{\text{sps}}^L(r, s)$  we approximate the improper integral in (8) with a finite sum and evaluate the norm of the matrix exponential  $\|e^{-\alpha L t}(r - s)\|$  at the points required by the appropriate discretization. An alternative notion of distance for graph-supported signals that is computationally more tractable comes in the form of the diffusion distance that we introduce in the next section.

#### IV. DIFFUSION DISTANCE

Given an arbitrary graph  $G = (V, E, W)$  with Laplacian  $L$ , an input vector norm  $\|\cdot\|$  and two signals  $r, s \in \mathbb{R}^n$  defined in the node space  $V$ , the diffusion distance  $d_{\text{diff}}^L(r, s)$  between  $r$  and  $s$  is given by

$$d_{\text{diff}}^L(r, s) := \left\| \int_0^{+\infty} e^{-t} e^{-\alpha L t} (r - s) dt \right\|, \quad (12)$$

with  $\alpha > 0$  corresponding to the diffusion constant in (5). As in the case of the superposition distance in (8), the diffusion distance incorporates the graph structure in determining the proximity between  $r$  and  $s$  through the solutions  $r(t)$  and  $s(t)$  of (5) for initial conditions  $r$  and  $s$  [cf. (6)]. The difference is that in the diffusion distance the input norm of the difference between  $r(t)$  and  $s(t)$  is taken *after* discounting and integration, whereas in the superposition distance the input norm is applied *before* discounting and integration. An interpretation in terms of heat diffusion is that the diffusion distance compares the total (discounted) energy that passes through each node. The superposition distance compares the energy difference at each point in time and integrates that difference over time. Both are reasonable choices. Whether the superposition or diffusion distance is preferable depends on the specific application.

A definite advantage of the diffusion distance is that the matrix integral in (12) can be resolved to obtain a closed solution that is more amenable to computation. To do so, notice that the primitive of the matrix exponential  $e^{-t} e^{-\alpha L t} = e^{-(I + \alpha L)t}$  is given by  $-(I + \alpha L)^{-1} e^{-(I + \alpha L)t}$  to conclude that (12) is equivalent to

$$d_{\text{diff}}^L(r, s) = \|(I + \alpha L)^{-1} (r - s)\|. \quad (13)$$

As in the case of the superposition distance of Section III a vector norm can be defined based on the same heat diffusion interpretation used to define the distance in (12). Therefore,

consider a given a graph with Laplacian  $L$  and a given input norm  $\|\cdot\|$  and define the diffusion norm of the vector  $v \in \mathbb{R}^n$  as

$$\|v\|_{\text{diff}}^L := \left\| \int_0^{+\infty} e^{-t} e^{-\alpha L t} v dt \right\| = \|(I + \alpha L)^{-1} v\|, \quad (14)$$

where the second equality follows from the same primitive expression used in (13).

The superposition distance is a proper metric and the superposition norm is a proper norm. We show first that  $\|\cdot\|_{\text{diff}}^L$  is a valid norm as we formally state next.

**Proposition 2** *The function  $\|\cdot\|_{\text{diff}}^L$  in (14) is a valid norm on  $\mathbb{R}^n$  for every Laplacian  $L$  and every input norm  $\|\cdot\|$ .*

**Proof:** To prove the validity of  $\|\cdot\|_{\text{diff}}^L$  we need to show positiveness, positive homogeneity and subadditivity; see Section II. Positive homogeneity follows directly from the positive homogeneity of the input norm, i.e. for any vector  $v \in \mathbb{R}^n$  and scalar  $\beta$  we have that

$$\begin{aligned} \|\beta v\|_{\text{diff}}^L &= \|(I + \alpha L)^{-1} \beta v\| \\ &= |\beta| \|(I + \alpha L)^{-1} v\| = |\beta| \|v\|_{\text{diff}}^L. \end{aligned} \quad (15)$$

In order to show subadditivity, pick arbitrary vectors  $v, w \in \mathbb{R}^n$  and use the subadditivity of the input norm  $\|\cdot\|$  to see that

$$\begin{aligned} \|v + w\|_{\text{diff}}^L &= \|(I + \alpha L)^{-1} (v + w)\| \\ &\leq \|(I + \alpha L)^{-1} v\| + \|(I + \alpha L)^{-1} w\| \\ &= \|v\|_{\text{diff}}^L + \|w\|_{\text{diff}}^L. \end{aligned} \quad (16)$$

Given the positiveness property of the input norm  $\|\cdot\|$ , to show positiveness of the diffusion norm  $\|\cdot\|_{\text{diff}}^L$  it is enough to show that  $(I + \alpha L)^{-1} v \neq \vec{0}$  for all vectors  $v \in \mathbb{R}^n$  different from the null vector. This is implied by the fact that  $(I + \alpha L)^{-1}$  is a positive definite matrix. To see why  $(I + \alpha L)^{-1}$  is positive definite, first notice that  $L$  is positive semidefinite as stated in Section II. Consequently,  $\alpha L$  is also positive semidefinite since  $\alpha > 0$  and  $I + \alpha L$  is positive definite since every eigenvalue of  $I + \alpha L$  is a unit greater than the corresponding eigenvalues of  $\alpha L$ , thus, strictly greater than 0. Finally, since inversion preserves positive definiteness, the proof is completed. ■

From Proposition 1 and Lemma 1 it follows directly that that the diffusion distance defined in (12) is a valid metric as we prove next.

**Corollary 2** *The function  $d_{\text{diff}}^L$  in (12) is a valid metric on  $\mathbb{R}^n$  for every Laplacian  $L$  and every input norm  $\|\cdot\|$ .*

**Proof:** Since  $d_{\text{diff}}^L(r, s) = \|r - s\|_{\text{diff}}^L$  for all vectors  $r, s \in \mathbb{R}^n$  and  $\|\cdot\|_{\text{diff}}^L$  is a well-defined norm [cf. Proposition 2], Lemma 1 implies that  $d_{\text{diff}}^L$  is a metric on  $\mathbb{R}^n$ . ■

As in the case of the superposition norm and distance, the diffusion norm and distance reduce to the input norm and its induced distance when the set edge is empty. In that case we have  $L = \mathbf{0}$  and it follows from the definitions in (14) and (12) that  $\|v\|_{\text{diff}}^L = \|v\|_{\text{diff}}^{\mathbf{0}} = \|v\|$  and that  $d_{\text{diff}}^L(r, s) = d_{\text{diff}}^{\mathbf{0}}(r, s) = \|r - s\|$ .

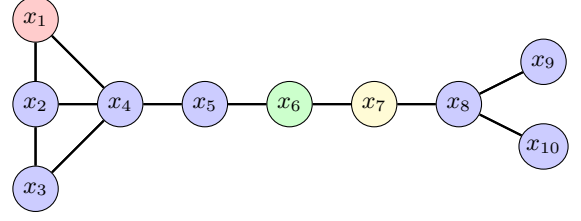


Fig. 1: Example of an underlying graph used to compute the superposition and diffusion distances. Three signals  $r$ ,  $g$  and  $y$  are compared taking a value of 1 in the red, green, and yellow nodes respectively, and zero everywhere else.

The superposition and diffusion distance differ in the order in which the input norm and time integral are applied. It is therefore reasonable to expect some relationship to hold between their values. In the following proposition we show that the diffusion distance is a lower bound for the value of the superposition distance.

**Proposition 3** *Given any graph  $G = (V, E, W)$  with Laplacian  $L$ , any two signals  $r, s \in \mathbb{R}^n$  defined in  $V$  and any input vector norm  $\|\cdot\|$ , the diffusion distance  $d_{\text{diff}}^L(r, s)$  defined in (12) is a lower bound on the superposition distance  $d_{\text{sps}}^L(r, s)$  defined in (8)*

$$d_{\text{sps}}^L(r, s) \geq d_{\text{diff}}^L(r, s). \quad (17)$$

**Proof:** Since the exponential  $e^{-t}$  in (8) is nonnegative, we may replace it with its absolute value to obtain

$$\begin{aligned} d_{\text{sps}}(r, s) &= \int_0^{+\infty} |e^{-t}| \|e^{-\alpha L t} (r - s)\| dt \\ &= \int_0^{+\infty} \|e^{-t} e^{-\alpha L t} (r - s)\| dt, \end{aligned} \quad (18)$$

where we used the positive homogeneity property of the input norm to write the second equality. Further using the subadditivity property of the input norm we may write

$$d_{\text{sps}}(r, s) \geq \left\| \int_0^{+\infty} e^{-t} e^{-\alpha L t} (r - s) dt \right\|. \quad (19)$$

The right hand side of (19) is the definition of the diffusion distance  $d_{\text{diff}}(r, s)$  in (12). Making this substitution in (19) yields (17). ■

For applications in which the superposition distance is more appropriate, the diffusion distance is still valuable because, as it follows from Proposition 3, it can be used as a lower bound on the superposition distance. This lower bound is useful because computing the diffusion distance is less expensive than computing the superposition distance.

#### A. Discussion

In order to illustrate the superposition and diffusion distances and their difference with the standard vector distances, consider the undirected graph in Figure 1 where the weight of each undirected edge is equal to 1. Define three different vectors supported in the node space and having exactly one component equal to 1 and the rest equal to 0. The vector  $r$  has

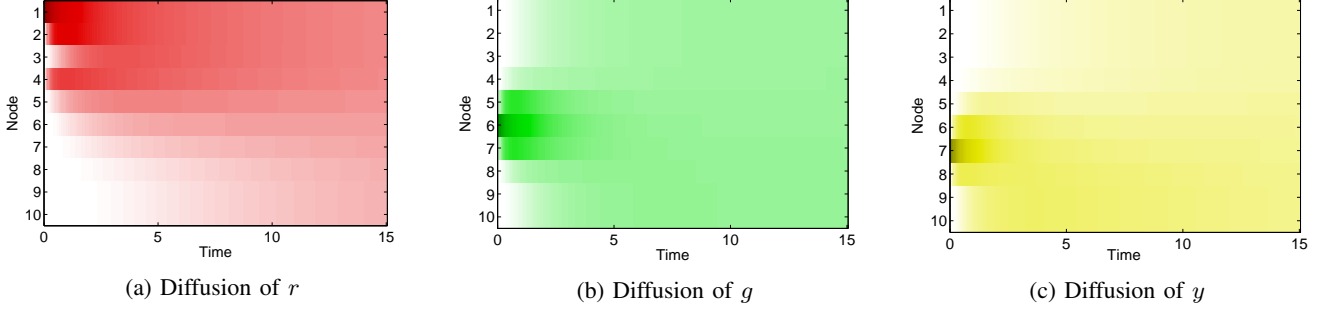


Fig. 2: Heat maps of the diffused signals for  $r$ ,  $g$ , and  $y$  as diffusion evolves for every node in the network in Figure 1. Darker colors represent stronger signals. The heat maps of  $g$  and  $y$  are more similar, entailing smaller diffusion and superposition distances.

its positive component for node  $x_1$ , colored in red, the vector  $g$  has its positive component for node  $x_6$ , colored in green, and the vector  $y$  has its positive component for node  $x_7$ , colored in yellow.

For the traditional vector metrics, the distances between each of the vectors  $r$ ,  $g$  and  $y$  is the same. In the case when, e.g., the  $\ell_2$  distance is used as input metric, we have that  $\|r - g\|_2 = \|g - y\|_2 = \|y - r\|_2 = \sqrt{2}$ . In the case of the  $\ell_1$  and  $\ell_\infty$  distances we have that  $\|r - g\|_1 = \|g - y\|_1 = \|y - r\|_1 = 2$  and  $\|r - g\|_\infty = \|g - y\|_\infty = \|y - r\|_\infty = 1$ . However, by observing the network in Figure 1, it is intuitive that signals  $g$  and  $y$  should be more alike than they are to  $r$  since they affect nodes that are closely related. E.g., if we think of the vectors  $r$ ,  $g$  and  $y$  as signaling faulty nodes in a communication network, it is evident that the impact of nodes  $x_6$  and  $x_7$  failing would disrupt the communication between the right and left components of the graph, whereas the failure of  $x_1$  would entail a different effect. This intuition is captured by the diffusion and superposition distances. Indeed, if we fix  $\alpha = 1$  and we use the  $\ell_2$  norm as input norm to the diffusion distance, we have that the distance between the vectors that signal faults at  $x_6$  and  $x_7$  are [cf. (13)]

$$d_{\text{diff}}^L(g, y) = \|(I + L)^{-1}(g - y)\|_2 = 0.418, \quad (20)$$

where  $L$  is the Laplacian of the graph in Figure 1. However, the diffusion distances from these green and yellow vectors to the red vector that signals a fault at node  $x_1$  are

$$\begin{aligned} d_{\text{diff}}^L(r, g) &= \|(I + L)^{-1}(r - g)\|_2 = 0.664, \\ d_{\text{diff}}^L(r, y) &= \|(I + L)^{-1}(r - y)\|_2 = 0.698. \end{aligned} \quad (21)$$

The distances in (21) are larger than the distance in (20) signaling the relative similarity of the  $g$  and  $y$  vectors with respect to the  $r$  vector. The differences are substantial – almost 60% increase –, thus allowing identification of  $g$  and  $y$  as somehow separate from  $r$ . Further observe that the distance between  $r$  and  $g$  is slightly smaller than the distance between  $r$  and  $y$ . This is as it should be, because node  $x_1$  is closer to node  $x_6$  than to node  $x_7$  in the underlying graph.

Repeating the exercise, but using the superposition distance instead [cf. (8)], we obtain that  $d_{\text{sps}}^L(r, g) = 0.701$ ,  $d_{\text{sps}}^L(r, y) = 0.742$ , and  $d_{\text{sps}}^L(g, y) = 0.456$ . Although the numbers are slightly different, the qualitative conclusions are the same as those obtained for the diffusion distance. We can tell that  $g$

and  $y$  are more like each other than they are to  $r$ , and we can tell that  $g$  is slightly closer to  $r$  than  $y$  is. Also note that the diffusion distances are smaller than the superposition distances between the corresponding pairs, i.e.,  $d_{\text{sps}}^L(r, g) \geq d_{\text{diff}}^L(r, g)$ ,  $d_{\text{sps}}^L(r, y) \geq d_{\text{diff}}^L(r, y)$ , and  $d_{\text{sps}}^L(g, y) \geq d_{\text{diff}}^L(g, y)$ . This is consistent with the result in Proposition 3.

To further illustrate the intuitive idea behind the diffusion and superposition distances, Figure 2 plots the evolution of the diffused signals  $r(t)$ ,  $g(t)$  and  $y(t)$  for each of the respective initial conditions  $r$ ,  $g$ , and  $y$ . At time  $t = 0$  each of the signals is concentrated at one specific node. The signals are, as a consequence, equally different to each other. At very long times, the signals are completely diffused and therefore indistinguishable. For intermediate times, the signal distributions across nodes for the green and yellow signals are more similar than between the green and red or yellow and red signals. This difference between the evolution of the diffused signals results in different values for the superposition and diffusion distances.

**Remark 1** Computation of the diffusion distance using the closed form expression in (13) requires the inversion of the  $n \times n$  identity plus Laplacian matrix followed by multiplication with the difference vector  $r - s$ . The cost of this computation is of order  $n^3$ , but is much smaller when the matrix  $L$  is sparse, as is typically the case. Further observe that most computations can be reused when computing multiple distances, because the vectors change, but the matrix inverse  $(I + \alpha L)^{-1}$  stays unchanged.

## V. STABILITY

The superposition and diffusion distances depend on the underlying graphs through their Laplacian  $L$ . It is therefore important to analyze how a perturbation of the underlying network impacts both distances. We prove in this section that these distances are well behaved with respect to perturbations of the underlying graph. I.e., we show that if the network perturbation is small, the change in the diffusion and superposition distances is also small. We quantify the network perturbation as the matrix  $p$ -norm of the difference between the Laplacians of the original and perturbed networks. We focus our analysis on the most frequently used norms where



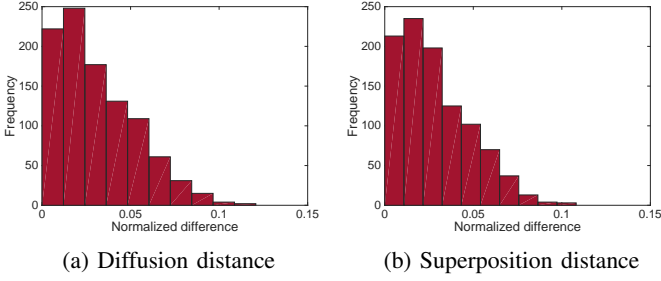


Fig. 3: Histogram of the absolute value of the normalized difference, i.e.  $|d^{L'}(g, r) - d^L(g, r)|/\|E\|_2$ , for the diffusion and superposition distances. For this particular network and perturbations, the difference is considerably lower than the theoretical upper bound of 2.

$p \in \{1, 2, \infty\}$ . We begin with a formal statement for the case of the superposition distance defined by (8).

**Theorem 1** *Given any graph with Laplacian  $L$ , an input  $\ell_p$  norm  $\|\cdot\|_p$  with  $p \in \{1, 2, \infty\}$ , and bounded signals  $s$  and  $r$  on the network with  $\|s\|_p \leq \gamma$  and  $\|r\|_p \leq \gamma$ , if we perturb the network such that the resulting Laplacian  $L' = L + E$  where the perturbation  $E$  is such that  $\|E\|_p \leq \epsilon\|L\|_p < 1$ , then*

$$|d_{sps}^{L'}(s, r) - d_{sps}^L(s, r)| \leq 2\gamma\|L\|_p\epsilon. \quad (22)$$

**Proof:** See Appendix A. ■

Theorem 1 guarantees that for any two vectors, the difference between their superposition distances computed based on different underlying graphs is bounded by a term which is bilinear in a bound on the magnitude of the input vectors  $\gamma$  and a bound on the difference between the Laplacians of both underlying graphs  $\|E\|_p \leq \epsilon\|L\|_p$ . This implies that vanishing perturbations on the underlying network have vanishing effects on the distance between two signals defined on the network.

Similarly to the case of the superposition distance, perturbations have limited effect on the diffusion metric defined in (12) as shown next.

**Theorem 2** *For the same setting described in Theorem 1, we have that*

$$|d_{diff}^{L'}(s, r) - d_{diff}^L(s, r)| \leq 2\gamma\|L\|_p\epsilon + o(\epsilon). \quad (23)$$

**Proof:** See Appendix B. ■

In contrast to Theorem 1, the bound in (23) contains higher order terms that depend on the magnitude of the perturbation. Hence, since the other terms of the bound in (23) tend to zero super linearly, we may divide (23) by  $\epsilon\|L\|_p$  and compute the limit as the perturbation vanishes

$$\lim_{\epsilon \rightarrow 0} \frac{|d_{diff}^{L'}(s, r) - d_{diff}^L(s, r)|}{\epsilon\|L\|_p} \leq 2\gamma, \quad (24)$$

which implies that for small perturbations the difference in diffusion distances grows linearly.

When constructing the underlying graph to compare signals in a real-world application, noisy information can be introduced. This means that the similarity weight between two nodes in the underlying graph contains inherent error. Theorems 1 and 2 show that the superposition and diffusion distances are impervious to these minor perturbations.

In order to illustrate the stability results presented, consider again the underlying network in Figure 1. We perturb this network by multiplying every edge weight – originally equal to 1 – by a random number uniformly picked from  $[0.95, 1.05]$  and then compute the diffusion and superposition distances between vectors  $r$  and  $g$  with the perturbed graph as underlying network. For these illustrations we pick the input norm to be  $\ell_2$  and observe that  $\gamma = 1$  given the definitions of  $r$  and  $g$ . In Figure 3 we plot histograms of the absolute value of the difference in the distances when using the original and the perturbed graphs as underlying networks normalized by the norm of the perturbation for 1000 repetitions of the experiment. From (22) we know that this value should be less than 2 for the superposition distance and from (24) we know this should also be the case for the diffusion distance for vanishing perturbations. Indeed, as can be seen from Figure 3, all perturbations are below the threshold of 2 by a considerable margin. This stability property is essential for the practical utility of the diffusion and superposition distances as seen in the next section.

**Remark 2** In Theorems 1 and 2 we focus our analysis on the input norms  $\|\cdot\|_p$  for  $p \in \{1, 2, \infty\}$  because these norms lead to the simple bounds in (22) and (23). The simplicity of these bounds is derived from the fact that  $\|e^{-Lt}\|_p \leq 1$  and  $\|(I + L)^{-1}\|_p \leq 1$  for the values of  $p$  previously mentioned. For other matrix norms satisfying (3) and (4), including all induced matrix norms, the equivalence of norms guarantees that bounds analogous to those in (22) and (23) must exist with more complex constant terms.

## VI. APPLICATIONS

We illustrate the advantages of the superposition and diffusion distances developed in Sections III and IV respectively through numerical experiments in both synthetic (Section VI-A) and real-world data (Section VI-B).

### A. Classification of synthetic signals on networks

The diffusion and superposition distances lead to better classification of signals on networks compared to traditional vector distances such as the Euclidean  $\ell_2$  metric. Consider the network presented in Figure 4a containing three clusters – blue, red, and green – where nodes within each cluster are highly connected and there exist few connections between nodes in different clusters. This network was generated randomly, where an undirected edge between a pair of nodes in the same cluster is formed with probability 0.4 and its weight is picked uniformly between 1 and 3. In addition, three edges were added with weight 1 between random pairs of nodes in different clusters. We consider three types of signals on this

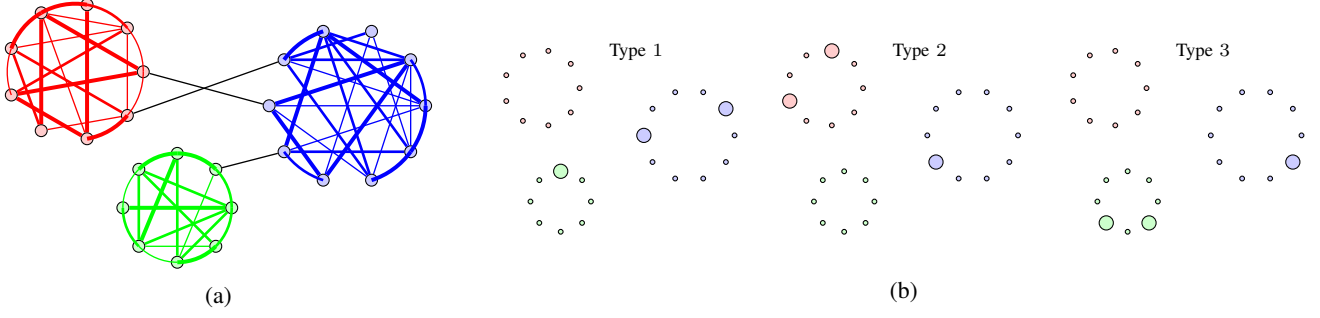


Fig. 4: (a) The three-cluster network on which signals to be classified are defined. The width of the links is proportional to the weights of the corresponding edges. (b) Sample signals for the three types considered. Type 1 signals have stronger presence in the blue cluster, type 2 in the red, and type 3 in the green cluster.

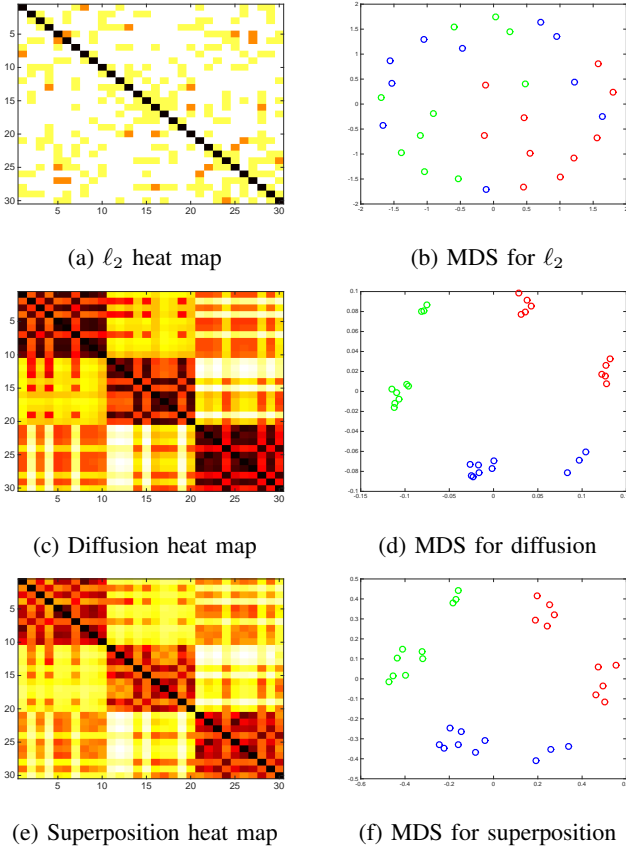


Fig. 5: Heat maps (left) and 2 dimensional MDS representations (right) for the metric spaces generated by the  $\ell_2$  (top), diffusion (middle) and superposition (bottom) distances. The diffusion and superposition metrics perfectly classify the signals into the three types while  $\ell_2$  does not reveal any clear classification.

network. The strength of all signals is equal to 1 on three nodes in the network and 0 on the remaining ones. Among the three nodes with value 1 for the first type of signals, two of them are randomly selected from the blue cluster and the remaining one is randomly chosen from the other clusters. Similarly, for the second type of signals, exactly two out of the three nodes with positive value belong to the red cluster and the remaining

one is chosen randomly between the blue and green clusters. Finally, the third type of signal has two positive values on the green cluster and the third value randomly chosen from the rest of the network. Sample signals for each type are illustrated in Figure 4b where positive signal values are denoted by larger nodes.

We generate ten signals of each type and measure the distance between them with the superposition, diffusion, and  $\ell_2$  metrics. For the superposition and diffusion metrics we use  $\ell_2$  as input norm and  $\alpha = 1$ . The use of each metric generates a different metric space with the thirty signals as the common underlying set of points. In order to illustrate these higher dimensional spaces, in Figure 5 (left) we present heat maps of the distance functions, where darker colors represent closer signals. It is clear that for the diffusion and superposition distances, three blocks containing ten points each appear along the diagonal in exact correspondence with the three types of signals. In contrast, the heat map corresponding to the  $\ell_2$  metric does not present any clear structure. To further illustrate these implications, in Figure 5 (right) we present 2D multi dimensional scaling (MDS) [32] representations of the three metric spaces. The points corresponding to type 1 signals are represented as blue circles, type 2 as red circles, and type 3 as green circles. The MDS representations for diffusion and superposition are fundamentally different from the one obtained for  $\ell_2$ . For the latter, the circles of different colors are spread almost randomly on the plane, with no clear clustering structure. For diffusion and superposition, in contrast, signals of different colors are clearly separated so that any clustering method is able to recover the original signal type.

### B. Ovarian cancer histology classification

We demonstrate that the diffusion distance can provide a better classification of histology subtypes for ovarian cancer patients than the traditional  $\ell_2$  metric. To do this, we consider 240 patients diagnosed with ovarian cancer corresponding to two different histology subtypes [33]: serous and endometrioid. Our objective is to recover the histology subtypes from patients' genetic profiles.

For each patient  $i$ , her genetic profile consists of a binary vector  $v_i \in \{0, 1\}^{2458}$  where, for each of the 2458 genes studied,  $v_i$  contains a 1 in position  $k$  if patient  $i$  presents



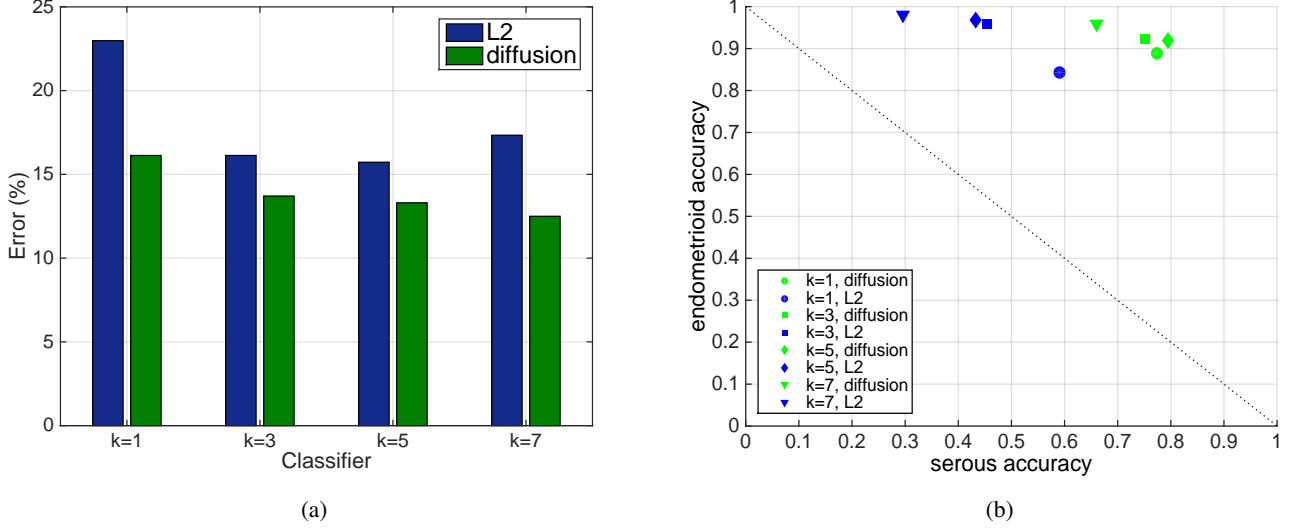


Fig. 6: Histology classification of ovarian cancer patients based on  $k$  nearest neighbors with respect to the  $\ell_2$  and diffusion distances of their genetic profile. (a) Blue bars denote the error when patients are classified using the  $\ell_2$  distance while the green bars denote the error when diffusion distance is used for different k-NN classifiers. The diffusion distance reduces the classification error consistently across classifiers. (b) Accuracy of serous subtype vs. endometrioid subtype. Classifiers using diffusion (green) are closer to the top right corner, i.e. perfect classification, than those using the  $\ell_2$  distance (blue).

a mutation in gene  $k$  and 0 otherwise. One way of building a metric in the space of 240 patients is by quantifying the distance between patients  $i$  and  $j$  as the  $\ell_2$  distance between their genetic profiles,

$$d_{\ell_2}(i, j) = \|v_i - v_j\|_2. \quad (25)$$

In this approach, every gene is considered orthogonal to each other and compared separately across patients. An alternative approach is to take into account the relational information across genes when comparing patients. In order to do so, we apply the diffusion distance on an underlying gene-to-gene network built based on publicly available data [34]. In order to build this network, we first extract the pairwise gene-gene interactions from [34] using the *NCI\_Nature* database. After normalization, every edge weight is contained between 0 and 1, which we interpret as a probability of interaction between genes. We assign to each path the probability obtained by multiplying the probabilities in the edges that form path. For every pair of genes in the network, we compute a similarity value between them corresponding to the maximum probability achievable by a path that links both genes. Finally, we apply normalization and thresholding operations to obtain the gene-to-gene network that we use in our experiments. Observe that the gene-to-gene network contains accepted relations between genes in humans in general and is not patient dependent, hence, it defines a common underlying network for all subjects being compared. Thus, denoting as  $L$  the Laplacian of the gene-to-gene network and using the  $\ell_2$  as input norm we compute the diffusion distances between patients  $i$  and  $j$  as [cf. (13)]

$$d_{\text{diff}}^L(i, j) = \|(I + \alpha L)^{-1}(v_i - v_j)\|_2, \quad (26)$$

where  $\alpha$  was set to 15, however, results are robust to this

particular choice. Given that in Section VI-A we obtained similar performance between the diffusion and superposition distances, combined with the fact that the latter is computationally expensive, we do not implement the superposition distance in this data set.

In order to evaluate the classification power of both approaches –  $\ell_2$  and diffusion distance – we perform 240-fold cross validation for a  $k$  nearest neighbors (k-NN) classifier. More precisely, for a particular patient, we look at the  $k$  nearest patients as given by the metric being evaluated and assign to this patient the most common cancer histology among the  $k$  nearest patients. We then compare the assigned histology with her real cancer histology and evaluate the accuracy of the classifier. Finally, we repeat this process for the 240 women considered and obtain a global classification accuracy of both approaches.

In Figure 6a we show the reduction in histology classification error when using the diffusion distance (26) compared to using the  $\ell_2$  distance (25) when comparing genetic profiles. The four groups of bars correspond to classifiers built using different numbers of neighbors  $k \in \{1, 3, 5, 7\}$ . Notice that the reduction in error is consistent across all classifiers analyzed with an average reduction of over 4% in the error rates, unveiling the value of incorporating the network information in the classification process.

To further analyze the obtained results, in Figure 6b we present the accuracy obtained for the serous subtype versus the accuracy obtained for the endometrioid subtype for different classifiers based on the diffusion (green) and  $\ell_2$  (blue) distances. Points on the top right corner of the plot are ideal, obtaining perfect classification for both subtypes. When using diffusion, accuracies shift towards the ideal position since the accuracies for the serous subtypes increase by 20% to 40%

whereas the accuracies for endometrioid subtypes decrease by less than 5%. Furthermore, among the 240 patients analyzed, there are 196 of them with endometrioid subtype and only 44 with serous subtype. Hence, a nearest neighbor classifier based on an uninformative distance would tend to have a high classification accuracy for the former but a low one for the latter. This is the case for the  $\ell_2$  metric. The diffusion distance, in contrast, by exploiting the gene-to-gene interaction can overcome this limitation.

## VII. FEATURE SPACE TRANSFORMATION

The diffusion distance in (13) between  $r, s \in \mathbb{R}^n$  can be interpreted as the input norm of the difference between two diffused vectors  $r_{\text{diff}}$  and  $s_{\text{diff}}$  also defined in  $\mathbb{R}^n$ , i.e.  $d_{\text{diff}}^L(r, s) = \|r_{\text{diff}} - s_{\text{diff}}\|$  where

$$r_{\text{diff}} = (I + \alpha L)^{-1} r, \quad (27)$$

and similarly for  $s_{\text{diff}}$ . Thus, diffusion can be seen as a transformation of the feature space for cases where there exists additional information about the relation between features. This relation is based on prior knowledge about the feature spaced instead of being data driven by particular observations. For example, for the genetic network in Section VI-B we have the additional information – independent of the set of patients – that there is interrelation between the function of some genes. Hence, we use these relations to define diffused mutations for each patient. However, apart from looking at the distance between the diffused signals – as proposed in Section IV and applied in Section VI – we can analyze the image of each signal under this transformation.

As an illustration, consider the well-known MNIST handwritten digit database [35]. Each observation consists of a square gray-scaled image of a handwritten digit with  $28 \times 28$  pixels. Consequently, we can think of each observation as a vector  $x \in \mathbb{R}^{784}$  where the value of each component corresponds to the intensity of the associated pixel. However, among these 784 features there are relations imposed by the lattice structure of the image. In particular, pixels found close in the image play a similar role in the specification of a handwritten digit. Thus, we build a lattice graph where each pixel is linked by an edge of unit weight to its contiguous pixels. If we denote by  $L$  the Laplacian of the lattice graph built, we may use (27) to obtain the diffused versions of different handwritten digits. In Figure 7 we present two different observations of the digit 3 as found in the MNIST database and after diffusion with  $\alpha = 0.8$ . From the figure, it is clear that diffusion smoothens imperfections of particular hand written instances, facilitating the comparison of diffused versions of the digits. E.g., the  $\ell_2$  distance between the two original images is 10.13 while the average distance between any two digits taken at random from the database is 10.19. However, after diffusion, the  $\ell_2$  distance between these two images is 6.41 while the average distance between any two diffused digits is 6.88, providing a better classification power. This motivates the use of diffusion as a preprocessing transformation for learning.

In this direction, we build two support vector machine (SVM) classifiers with a radial basis function kernel [36] to

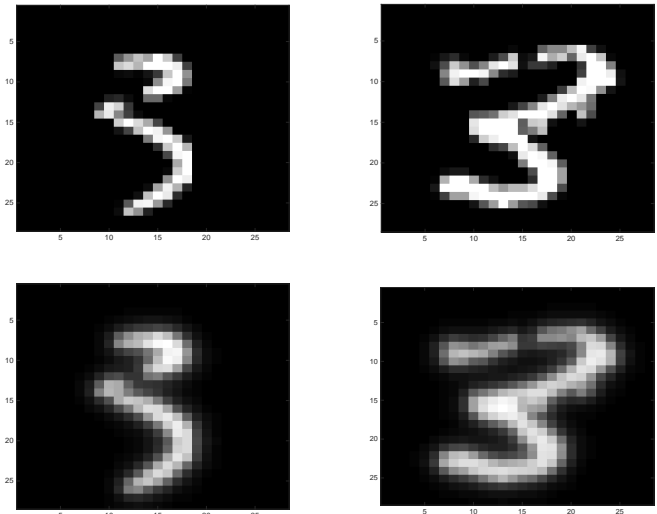


Fig. 7: Two samples of the handwritten digit 3 (top row) and their corresponding diffused versions (bottom row). The imperfections of the original handwritten samples are smoothed by diffusion making it easier to compare the diffused versions of the digits.

recognize handwritten digits. The first classifier is trained on the original space  $\mathbb{R}^{784}$  where each feature corresponds to the intensity of one particular pixel whereas the second one is also built on  $\mathbb{R}^{784}$  but after transforming the space using diffusion. In Figure 8 we present the error rates for SVM classification between subsets of digits which are hard to distinguish, such as 3 and 5. For these experiments, we choose 800 random samples of the digits being analyzed from the MNIST database and partition the sampled data into two halves corresponding to the training and testing data. Within the training data we perform 5-fold cross validation to select the best combinations of the penalty parameter for the error term in the SVM objective function and the spreading parameter in the radial basis function kernel. We then train the SVM using the entire training set with the best parameter combinations and compute the accuracy of using the trained model to classify the testing data. From the figure it is immediate that the diffusion transformation reduces the classification error, e.g. when distinguishing between 3, 5, 8, and 9, the error is reduced from 4.5% to 2%, and when distinguishing between 1, 2, and 7 the error is reduced from 1.75% to perfect attribution. Similarly, we compare the accuracy of both approaches when training a multi-class classifier to categorize among the ten possible digits. We run this experiment for a sample size of 2000 digits equally distributed across the ten possible digits of which 80% is considered training data and the rest testing data. We follow the same training procedure described for the classification of subsets of digits. The total error obtained by the original approach is 5.75% whereas by using diffusion to preprocess the data we reduce this error to 4.25%, i.e. a 26% reduction of the error rate.

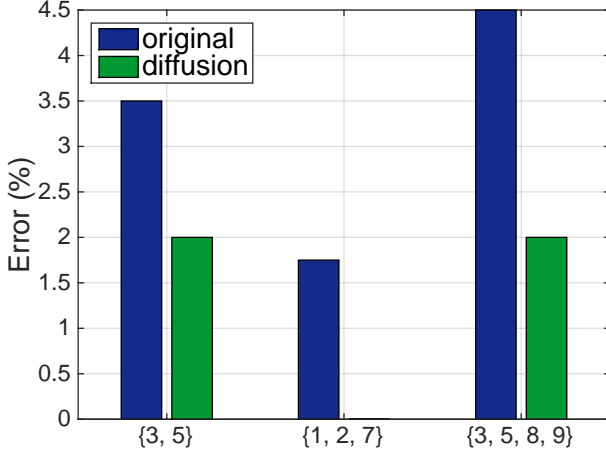


Fig. 8: Error rates for a two class, a three class and a four class classification of written digits given by a SVM trained in the original and the diffused data set. The error is reduced by diffusion in the three cases.

### VIII. CONCLUSION

The superposition and diffusion distances, as metrics to compare signals in networks, were introduced. Both metrics rely on the temporal heat map induced by the diffusion of signals across the network. The superposition distance quantifies the instantaneous difference between the diffused signals while the diffusion distance evaluates the accumulated effect across time. Both distances were shown to be stable with respect to perturbations in the underlying network, however, due to its closed form, the diffusion distance was found to be more suitable for implementation. We showed how both distances can be used to obtain a better classification of signals in networks both in synthetic settings as well as in a real-world classification of cancer histologies. Finally, we reinterpreted diffusion as a transformation of the feature space which can be used as a preprocessing step in learning, and illustrated its utility by classifying handwritten digits.

### APPENDIX A PROOF OF THEOREM 1

The following lemma is central to the proof of Theorem 1.

**Lemma 2** *Given the Laplacian  $L$  for some undirected network, the matrix exponential of nonpositive multiples of the Laplacian  $e^{-\tau L}$  with  $\tau \geq 0$  is a doubly stochastic matrix.*

**Proof:** Since  $L = D - A$ , all off-diagonal components of  $L$  are nonpositive, therefore  $-L$  and  $-\tau L$  are Metzler matrices. Since the exponentials of Metzler matrices are nonnegative [37, Theorem 8.2], we are guaranteed that all elements of  $e^{-\tau L}$  are nonnegative. From the power series of matrix exponentials, we have

$$e^{-\tau L} = \sum_{k=0}^{\infty} \frac{1}{k!} (-\tau L)^k = I - \tau L + \frac{\tau^2 L^2}{2} - \frac{\tau^3 L^3}{3!} + \dots \quad (28)$$

If we are able to show that all rows and columns of  $L^k$  add up to 0 for any integer  $k \geq 1$ , then we know that all rows and columns of  $\sum_{k=1}^{\infty} (-\tau L)^k / k!$  also add up to 0. Therefore, when we add the identity matrix to this summation to obtain the exponential  $e^{-\tau L}$  as in (28) we are guaranteed that the rows and columns sum up to 1. Combining this with the non negativity of  $e^{-\tau L}$  implies doubly stochasticity, as wanted. We now prove that all rows and columns of  $L^k$  indeed add up to 0 for any integer  $k \geq 1$ . First notice that for  $k = 1$  this is immediate since the rows and columns of the Laplacian sum up to 0 by definition. Now, consider an arbitrary matrix  $B = C L$  obtained by left multiplying  $L$  by another matrix  $C$ . Then, the sum of any row of  $B$  is given by

$$\sum_j B_{ij} = \sum_j \sum_m C_{im} L_{mj} = \sum_m C_{im} \sum_j L_{mj} = 0, \quad (29)$$

where the last equality follows from the fact that  $\sum_j L_{mj} = 0$  for any  $m$ , i.e. all rows of the Laplacian sum up to 0. Similarly, we can show that the columns of a matrix  $B = L C$  obtained by right multiplying the Laplacian by another matrix  $C$  sum up to 0. Finally, for any power  $k$ , the matrix  $L^k = L^{k-1} L = L L^{k-1}$  can be obtained by both right or left multiplying  $L^{k-1}$  by the Laplacian  $L$ , thus all rows and columns of  $L^k$  sum up to 0 for all  $k \geq 1$ . ■

We now use Lemma 2 to show Theorem 1.

**Proof of Theorem 1:** Given the definition of  $L'$ , from (8) we have that

$$d_{\text{sps}}^{L'}(s, r) = \int_0^{\infty} e^{-t} \left\| e^{-(L+E)t} (s - r) \right\|_p dt, \quad (30)$$

where without loss of generality we assume  $\alpha = 1$ . If  $\alpha \neq 1$ , then  $\alpha L'$  defines a Laplacian and we can think of the distance  $d_{\text{sps}}^{\alpha L'}(s, r)$  where the new  $\alpha$  parameter is equal to 1. If we focus on the input norm  $\|\cdot\|_p$  inside the integral in (30), we may add and subtract  $e^{-Lt}(s - r)$  to obtain

$$\begin{aligned} \left\| e^{-(L+E)t} (s - r) \right\|_p &= \left\| (e^{-(L+E)t} - e^{-Lt}) (s - r) + e^{-Lt} (s - r) \right\|_p \\ &\leq \left\| (e^{-(L+E)t} - e^{-Lt}) (s - r) \right\|_p + \left\| e^{-Lt} (s - r) \right\|_p, \end{aligned} \quad (31)$$

where we used the subadditivity property of the input norm. To further bound the first term on the right hand side of (31) we apply the compatibility property of  $p$ -norms (4) followed by the subadditivity property to obtain that

$$\begin{aligned} \left\| (e^{-(L+E)t} - e^{-Lt}) (s - r) \right\|_p &\leq \left\| e^{-(L+E)t} - e^{-Lt} \right\|_p \left\| (s - r) \right\|_p \\ &\leq \left\| e^{-(L+E)t} - e^{-Lt} \right\|_p (\|s\|_p + \|r\|_p). \end{aligned} \quad (32)$$

In order to bound the first term on the right hand side of (32), we use a well-known result in matrix exponential analysis [38], [39] that allows us to write the difference of matrix exponentials in terms of an integral,

$$\begin{aligned} \left\| e^{-(L+E)t} - e^{-Lt} \right\|_p &= \left\| \int_0^t e^{-L(t-\tau)} E e^{-(L+E)\tau} d\tau \right\|_p \\ &\leq \int_0^t \left\| e^{-L(t-\tau)} E e^{-(L+E)\tau} \right\|_p d\tau \\ &\leq \|E\|_p \int_0^t \left\| e^{-L(t-\tau)} \right\|_p \left\| e^{-(L+E)\tau} \right\|_p d\tau, \end{aligned} \quad (33)$$

where the first inequality follows from subadditivity of the input  $p$ -norm and the second one from submultiplicativity (3).

We now bound each of the three terms on the right hand side of (33). For the first term,  $\|E\|_p \leq \epsilon \|L\|_p$  by assumption. From Lemma 2, the doubly stochasticity of  $e^{-L(t-\tau)}$  implies that  $\|e^{-L(t-\tau)}\|_1 = \|e^{-L(t-\tau)}\|_\infty = 1$ . For  $p = 2$ ,  $-L$  being negative semi-definite with largest eigenvalue at 0 implies that the largest eigenvalue of  $e^{-L(t-\tau)}$  is equal to 1 and hence  $\|e^{-L(t-\tau)}\|_2 = 1$ . For the term  $\|e^{-(L+E)\tau}\|_p$ , notice that  $L + E = L'$  is in itself a Laplacian, meaning that we can follow the aforementioned argument and upper bound this term by 1. Substituting these bounds in (33) and solving the integral yields

$$\left\| e^{-(L+E)t} - e^{-Lt} \right\|_p \leq \epsilon \|L\|_p t. \quad (34)$$

Further substitution in (32) combined with the fact that  $\|s\|_p \leq \gamma$  and  $\|r\|_p \leq \gamma$ , results in

$$\left\| \left( e^{-(L+E)t} - e^{-Lt} \right) (s - r) \right\|_p \leq 2\gamma\epsilon \|L\|_p t. \quad (35)$$

By substituting this result in (31) and inputting the resultant inequality in the integral in (30) we conclude that

$$d_{\text{sps}}^{L'}(s, r) \leq \int_0^\infty t e^{-t} 2\gamma\epsilon \|L\|_p dt + \int_0^\infty e^{-t} \|e^{-Lt}(s - r)\|_p dt. \quad (36)$$

Notice that the rightmost summand in (36) is exactly equal to  $d_{\text{sps}}^L(r, s)$  [cf. (8)]. Thus, solving the integral in the first summand we get that

$$d_{\text{sps}}^{L'}(s, r) - d_{\text{sps}}^L(s, r) \leq 2\gamma\epsilon \|L\|_p. \quad (37)$$

Following the same methodology but starting from the definition of  $d_{\text{sps}}^L(s, r)$ , it can be shown that

$$d_{\text{sps}}^L(s, r) - d_{\text{sps}}^{L'}(s, r) \leq 2\gamma\epsilon \|L\|_p. \quad (38)$$

Finally, by combining (37) and (38), we obtain (22), concluding the proof. ■

## APPENDIX B PROOF OF THEOREM 2

In the proof of Theorem 2 we use two lemmas. The first one is similar to Lemma 2 and shows that  $(I + L)^{-1}$  is doubly stochastic.

**Lemma 3** *Given the Laplacian  $L$  for some undirected network, the inverse of the Laplacian plus identity matrix  $(I + L)^{-1}$  is a doubly stochastic matrix.*

**Proof:** Since all the off-diagonal entries of  $I + L$  are less than or equal to zero,  $I + L$  is a  $Z$ -matrix [40]. Moreover, due to the fact that all eigenvalues of  $I + L$  have positive real parts,  $I + L$  is an  $M$ -matrix. Since the inverse of an  $M$ -matrix is elementwise nonnegative [41],  $(I + L)^{-1}$  is a nonnegative matrix. Thus, to show doubly stochasticity, we only need to prove that all rows and columns of  $(I + L)^{-1}$  add up to 1.

Denote entries in  $(I + L)$  as  $l_{ij}$  and in  $(I + L)^{-1}$  as  $a_{ij}$ , from  $(I + L)^{-1}(I + L) = I$ , we know that for any  $i$ ,

$$\sum_k a_{ik} l_{ki} = I_{ii} = 1, \quad (39)$$

$$\sum_k a_{ik} l_{kj} = I_{ij} = 0, \text{ for all } j \neq i. \quad (40)$$

Summing (40) over all  $j$  yields

$$\sum_j \left( \sum_k a_{ik} l_{kj} \right) = \sum_k a_{ik} \left( \sum_j l_{kj} \right) = 1. \quad (41)$$

Since  $\sum_j l_{kj} = 1$  for any  $k$  from the definition of the matrix  $(I + L)$ , we know that  $\sum_k a_{ik} = 1$  implying that the summation of any rows of  $(I + L)^{-1}$  is 1. Similarly,  $(I + L)(I + L)^{-1} = I$  induces that the summation of all columns of  $(I + L)^{-1}$  is 1, concluding the proof. ■

The second lemma is a statement about the stability of inverse matrices.

**Lemma 4** *If  $A$  is nonsingular and  $\|A^{-1}E\|_p < 1$ , then  $A + E$  is nonsingular and it is guaranteed that*

$$\|(A + E)^{-1} - A^{-1}\|_p \leq \frac{\|E\|_p \|A^{-1}\|_p^2}{1 - \|A^{-1}E\|_p}. \quad (42)$$

**Proof:** See [31, Theorem 2.3.4]. ■

We now use Lemmas 3 and 4 to show Theorem 2.

**Proof of Theorem 2:** Given the definition of  $L'$ , from (13) we have that

$$d_{\text{diff}}^{L'}(s, r) = \|(I + L + E)^{-1}(s - r)\|_p. \quad (43)$$

As in the proof of Theorem 1, we can assume that  $\alpha = 1$  without loss of generality. Subtracting and adding  $(I + L)^{-1}(s - r)$  from (43) and applying the subadditivity property of the  $p$ -norm implies

$$d_{\text{diff}}^{L'}(s, r) \leq \|((I + L + E)^{-1} - (I + L)^{-1})(s - r)\|_p + \|(I + L)^{-1}(s - r)\|_p, \quad (44)$$

where the second term in the sum is exactly  $d_{\text{diff}}^L(s, r)$  [cf. (13)]. Therefore we may write

$$d_{\text{diff}}^{L'}(s, r) - d_{\text{diff}}^L(s, r) \leq \|((I + L + E)^{-1} - (I + L)^{-1})(s - r)\|_p. \quad (45)$$

By applying compatibility of  $p$ -norms (4) followed by the subadditivity property we obtain that

$$\begin{aligned} d_{\text{diff}}^{L'}(s, r) - d_{\text{diff}}^L(s, r) & \leq \|((I + L + E)^{-1} - (I + L)^{-1})\|_p \|s - r\|_p \\ & \leq \|((I + L + E)^{-1} - (I + L)^{-1})\|_p (\|s\|_p + \|r\|_p) \end{aligned} \quad (46)$$

Given that  $I + L$  is nonsingular we have to show that  $\|(I + L)^{-1}E\|_p < 1$  in order to be able to apply Lemma 4 with  $A = (I + L)$  and further bound (46).

Due to doubly stochasticity [cf. Lemma 3], we have that  $\|(I + L)^{-1}\|_1 = \|(I + L)^{-1}\|_\infty = 1$ . Moreover,  $\|(I + L)^{-1}\|_2 = 1$  comes from the fact that the smallest eigenvalue

of  $(I + L)$  and hence the largest eigenvalue of  $(I + L)^{-1}$  is equal to 1. Consequently, we may write

$$\|(I + L)^{-1}E\|_p \leq \|(I + L)^{-1}\|_p \|E\|_p < 1, \quad (47)$$

for  $p \in \{1, 2, \infty\}$ , as wanted, where the first inequality follows from submultiplicativity (3). Hence, applying Lemma 4 with  $A = (I + L)$  yields

$$\|(I + L + E)^{-1} - (I + L)^{-1}\|_p \leq \frac{\|E\|_p \|(I + L)^{-1}\|_p^2}{1 - \|(I + L)^{-1}E\|_p}. \quad (48)$$

Recalling that  $\|(I + L)^{-1}\|_p = 1$  for any  $p \in \{1, 2, \infty\}$  allows us to further bound (48) to obtain

$$\|(I + L + E)^{-1} - (I + L)^{-1}\|_p \leq \frac{\|E\|_p}{1 - \|E\|_p} \leq \frac{\epsilon \|L\|_p}{1 - \epsilon \|L\|_p}, \quad (49)$$

where we used that  $\|E\|_p \leq \epsilon \|L\|_p < 1$  for the last inequality.

Utilizing the Taylor series of  $1/(1 - \epsilon \|L\|_p)$  and substituting (49) into (46) combined with the fact that  $\|s\|_p \leq \gamma$  and  $\|r\|_p \leq \gamma$  we have that

$$d_{\text{diff}}^{L'}(s, r) - d_{\text{diff}}^L(s, r) \leq \sum_{n=1}^{\infty} 2\gamma(\epsilon \|L\|_p)^n = 2\gamma \|L\|_p \epsilon + o(\epsilon). \quad (50)$$

In a similar manner but starting from the definition of  $d_{\text{diff}}^L(s, r)$ , it can be shown that

$$d_{\text{diff}}^L(s, r) - d_{\text{diff}}^{L'}(s, r) \leq 2\gamma \|L\|_p \epsilon + o(\epsilon). \quad (51)$$

Finally, by combining (50) and (51), we obtain (23) and the proof concludes. ■

## REFERENCES

- [1] D. Bu, Y. Zhao, L. Cai, H. Xue, X. Zhu, H. Lu, J. Zhang, S. Sun, L. Ling, and N. Zhang, "Topological structure analysis of the protein-protein interaction network in budding yeast," *Nucleic acids research*, vol. 31, no. 9, pp. 2443–2450, 2003.
- [2] E. Lieberman, C. Hauert, and M. Nowak, "Evolutionary dynamics on graphs," *Nature*, vol. 433, no. 7023, pp. 312–316, 2005.
- [3] M. E. J. Newman, "Finding community structure in networks using the eigenvectors of matrices," *Phys. Rev. E*, vol. 74, p. 036104, 2006.
- [4] J. Kleinberg, "Authoritative sources in a hyperlinked environment," *J. ACM*, vol. 46, no. 5, pp. 604–632, Sep. 1999.
- [5] —, "Complex networks and decentralized search algorithms," in *Proceedings of the International Congress of Mathematicians (ICM)*, vol. 3, 2006, pp. 1019–1044.
- [6] D. Kempe and F. McSherry, "A decentralized algorithm for spectral analysis," in *Proceedings of the Thirty-sixth Annual ACM Symposium on Theory of Computing*, New York, NY, USA, 2004, pp. 561–568.
- [7] N. Lynch, *Distributed Algorithms*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1996.
- [8] J. Noble and D. Boukerroui, "Ultrasound image segmentation: a survey," *Medical Imaging, IEEE Transactions on*, vol. 25, no. 8, pp. 987–1010, Aug 2006.
- [9] B. Miller, N. Bliss, and P. Wolfe, "Toward signal processing theory for graphs and non-euclidean data," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, March 2010, pp. 5414–5417.
- [10] D. Shuman, S. Narang, P. Frossard, A. Ortega, and P. Vandergheynst, "The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains," *Signal Processing Magazine, IEEE*, vol. 30, no. 3, pp. 83–98, May 2013.
- [11] A. Sandryhaila and J. Moura, "Discrete signal processing on graphs," *arXiv preprint arXiv:1210.4752*, 2012.
- [12] S. Narang and A. Ortega, "Downsampling graphs using spectral theory," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, May 2011, pp. 4208–4211.
- [13] R. Mittler, S. Vanderauwera, M. Gollery, and F. V. Breusegem, "Reactive oxygen gene network of plants," *Trends in Plant Science*, vol. 9, no. 10, pp. 490 – 498, 2004.
- [14] O. Sporns, *Networks of the Brain*. MIT press, 2011.
- [15] A. Luikov, *Analytical heat diffusion theory*. Academic press, New York, 1968.
- [16] E. Eckert and R. Drake, *Analysis of heat and mass transfer*. Hemisphere Publishing; New York, NY, 1987.
- [17] R. I. Kondor and J. Lafferty, "Diffusion kernels on graphs and other discrete input spaces," in *ICML*, vol. 2, 2002, pp. 315–322.
- [18] P. Carrington, J. Scott, and S. Wasserman, *Models and methods in social network analysis*. Cambridge University Press, 2005, vol. 28.
- [19] M. Freidlin and A. D. Wentzell, "Diffusion processes on graphs and the averaging principle," *The Annals of Probability*, vol. 21, no. 4, pp. pp. 2215–2245, 1993.
- [20] A. Szlam, M. Maggioni, and R. Coifman, "Regularization on graphs with function-adapted diffusion processes," *J. Mach. Learn. Res.*, vol. 9, pp. 1711–1739, 2008.
- [21] A. Smola and R. Kondor, "Kernels and regularization on graphs," in *Learning Theory and Kernel Machines*, ser. Lecture Notes in Computer Science, B. Scholkopf and M. Warmuth, Eds. Springer Berlin Heidelberg, 2003, vol. 2777, pp. 144–158.
- [22] W. Ren, R. W. Beard, and E. M. Atkins, "A survey of consensus problems in multi-agent coordination," *American Control Conference*, 2005.
- [23] J. A. F. R. Olfati-Saber and R. M. Murray, "Consensus and cooperation in networked multi-agent systems," *Proceedings of the IEEE*, vol. 95, no. 1, pp. 215–233, 2007.
- [24] W. Ren, "Consensus based formation control strategies for multi-vehicle systems," *American Control Conference*, 2006.
- [25] H. G. Tanner, A. Jadbabaie, and G. J. Pappas, "Stable flocking of mobile agents, part i: Fixed topology," *Conference on Decision and Control*, 2003.
- [26] M. H. DeGroot, "Reaching a consensus," *Journal of the American Statistical Association*, 1974.
- [27] J. C. Dittmer, "Consensus formation under bounded confidence," *Non-linear Analysis*, vol. 47, 2001.
- [28] S. Segarra and A. Ribeiro, "Hierarchical clustering and consensus in trust networks," in *Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP), 2013 IEEE 5th International Workshop on*, Dec 2013, pp. 85–88.
- [29] F. Chung, *Spectral graph theory*. American Mathematical Soc., 1997, vol. 92.
- [30] D. Burago, Y. Burago, and S. Ivanov, *A course in metric geometry*. American Mathematical Society Providence, 2001, vol. 33.
- [31] G. Golub and C. V. Loan, *Matrix Computations*. Johns Hopkins University Press, 1989.
- [32] M. A. A. Cox and T. F. Cox, "Multidimensional scaling," in *Handbook of Data Visualization*, ser. Springer Handbooks Comp.Statistics. Springer Berlin Heidelberg, 2008, pp. 315–347.
- [33] M. Hofree, J. Shen, H. Carter, A. Gross, and T. Ideker, "Network-based stratification of tumor mutations," *Nature methods*, 2013.
- [34] E. G. Cerami, B. E. Gross, E. Demir, I. Rodchenkov, O. Babur, N. Anwar, N. Schultz, G. Bader, and C. Sander, "Pathway commons, a web resource for biological pathway data," *Nucleic Acids Research*, vol. 39, no. suppl 1, pp. D685–D690, 2011. [Online]. Available: [http://nar.oxfordjournals.org/content/39/suppl\\_1/D685.abstract](http://nar.oxfordjournals.org/content/39/suppl_1/D685.abstract)
- [35] Y. Lecun and C. Cortes, "The MNIST database of handwritten digits." [Online]. Available: <http://yann.lecun.com/exdb/mnist/>
- [36] B. Scholkopf, S. Kah-Kay, C. Burges, F. Girosi, P. Niyogi, T. Poggio, and V. Vapnik, "Comparing support vector machines with gaussian kernels to radial basis function classifiers," *Signal Processing, IEEE Transactions on*, vol. 45, no. 11, pp. 2758–2765, Nov 1997.
- [37] R. Varga, "Matrix iterative analysis," *Springer series in computational mathematics*, 2000.
- [38] R. Bellman, *Introduction to matrix analysis*. SIAM, 1970, vol. 960.
- [39] C. V. Loan, "The sensitivity of the matrix exponential," *SIAM Journal on Numerical Analysis*, vol. 14, no. 6, pp. 971–981, 1977.
- [40] D. M. Young, *Iterative solution of large linear systems*. New York, Academic Press, 1971.
- [41] T. Fujimoto and R. Ranade, "Two characterizations of inverse-positive matrices: the hawkins-simon condition and the le chatelier-braun principle," *Electronic Journal of Linear Algebra*, vol. 11, pp. 59–65, 2004.